CLAIMS

1. Apparatus for identifying topics of document data, the apparatus comprising:

a word ranker operable to rank words that are present in or representative of the content of the document data;

a co-occurrence ranker operable to rank co-occurrences of words that are present in or representative of the content of the document data;

a phrase ranker operable to rank phrases in the document data;

a words selector operable to select the highest ranking words;

a co-occurrence identifier operable to identify which of the highest ranking co-occurrences contain at least one of the highest ranking words;

a phrase identifier operable to identify the phrases containing at least one word from the identified co-occurrences;

a phrase selector operable to select the highest ranking one or ones of the identified phrases as the topic or topics of the document data; and

an outputter operable to output data relating to the selected topics.

2. Apparatus according to claim 1, wherein the words selector is arranged to select as the highest ranking words a predetermined number of the highest ranking words, a number of the highest ranking words that represents a predetermined percentage of the words in the document data, or a number of the highest ranking words that represents a predetermined percentage of the number of ranked words.

3. Apparatus according to claim 1, wherein the co-occurrence identifier is arranged to select as the highest ranking co-occurrences a predetermined number of co-occurrences, a number of the highest ranking co-occurrences that represents a predetermined percentage of the co-occurrences in the document data, or a number of the highest ranking co-occurrences that represents a predetermined percentage of the number of ranked co-occurrences.

4. Apparatus according to claim 1, wherein the phrase selector is arranged to select as the highest ranking identified phrases a predetermined number of the identified phrases, a number of the highest ranking identified phrases that represents a predetermined percentage of the identified phrases in

the document data, or a number of the highest ranking identified phrases that represents a predetermined percentage of the number of ranked phrases.

5. Apparatus according to claim 1, wherein the phrase identifier is arranged to identify phrases by concatenating consecutive nouns, concatenating consecutive proper nouns, and concatenating consecutive adjectives with a final noun.

6. Apparatus according to claim 1, wherein at least one of the word ranker, co-occurrence ranker, and phrase ranker is arranged to weight the items to be ranked in accordance with their position in the document data.

7. Apparatus according to claim 1, further comprising a co-occurrence determiner operable to determine word co-occurrences in the document data by identifying as co-occurrences word combinations comprising words in particular grammatical categories.

8. Apparatus according to claim 1, further comprising a co-occurrence determiner operable to determine word co-occurrences in the document data by

identifying as co-occurrences at least some of the following combinations: noun and verb; noun and noun; noun and proper noun; verb and proper noun; and proper noun and proper noun.

9. Apparatus according to claim 7, wherein the co-occurrence determiner is arranged to ignore the order of the words in the word combinations.

10. Apparatus according to claim 1, wherein the co-occurrence ranker is arranged to rank significant co-occurrences and the apparatus further comprises a co-occurrence determiner operable to determine word co-occurrences in the document data by identifying as co-occurrences word combinations comprising words in particular grammatical categories and a significance calculator operable to calculate a significance measure for the identified co-occurrences.

11. Apparatus according to claim 1, wherein the co-occurrence ranker is arranged to rank significant co-occurrences and the apparatus further comprises a co-occurrence determiner operable to determine word co-occurrences in the document data by identifying as co-occurrences at least some of the following

combinations: noun and verb; noun and noun; noun and proper noun; verb and proper noun; and proper noun and proper noun, and a significance calculator operable to calculate a significance measure for the identified co-occurrences.

12. Apparatus according to claim 1, further comprising: a text splitter operable to split the document data into text segments; and a classifier operable to classify the selected topics according to the distribution in the text segments so as to define main and subsidiary topics in the document data, wherein the outputter is arranged to output data relating to the classified topics.

13. Apparatus according to claim 12, wherein the classifier is arranged to determine that a topic is a main topic if the topic occurs in a predetermined percentage of the text segments and to classify any topic not meeting this requirement as a subsidiary or lesser topic.

14. Apparatus according to claim 12, wherein the classifier is arranged to weight a topic in accordance

with the position in the document data of the text segment containing the topic.

15. Apparatus according to claim 12, wherein the classifier is arranged to weight a topic in accordance with the position in the document data of the text segments containing the topic so that a topic occurring in at least one of the first and last text segment of document data representing a document is given a higher weighting than topics occurring in the other text segments.

16. Apparatus according to claim 12, further comprising a topic hierarchy identifier operable to identify a topic as being a child or subsidiary topic of another topic when text portions in which that subsidiary topic occurs represent a sub-set of the text portions in which the said other topic occurs, wherein the outputter is arranged to output data relating to the identified topic hierarchy.

17. Apparatus according to claim 12, further comprising a topic hierarchy identifier operable to identify a topic as being a child or subsidiary topic of another topic when the text segments in which that

subsidiary topic occurs represent a sub-set of the text segments in which the said other topic occurs, wherein the outputter is arranged to output data relating to the identified topic hierarchy.

18. Apparatus according to claim 1, further comprising a summary provider operable to provide summary data on the basis of the selected topics, wherein the outputter is arranged to output the summary data.

19. Apparatus according to claim 18, wherein the summary provider comprises a sentence selector operable to select sentences for use in the summary data.

20. Apparatus according to claim 19, wherein the sentence selector comprises:

a topic weight assigner operable to assign weights to the topics;

a sentence weight assigner operable to assign weights to sentences in the document data;

a scorer operable to score the sentences by summing the assigned topic and sentence weights; and

a selector operable to select the sentence or sentences having the highest score or scores for the summary.

21. Apparatus according to claim 19, wherein the sentence selector comprises:

a topic weight assigner operable to assign weights to the topics;

a sentence weight assigner operable to assign weights to sentences in the document data;

a scorer operable to sccore the sentences by summing the assigned topic and sentence weights;

a selector operable to select the sentence or sentences having the highest score or scores;

a topic weight adjuster operable to relatively reduce the weight allocated to the topic or topics in the selected sentence or sentences; and

a controller operable to cause the scorer, selector and topic weight adjuster to repeat the above operations until a predetermined number of sentences has been selected for the summary from the document data.

22. Apparatus according to claim 21, wherein the topic weight adjuster is arranged to set to zero the

weight of any topic in the selected sentence or sentences.

23. Apparatus according to claim 19, further comprising:

a chunk identifier operable to identify in sentences selected for a summary chunks that do not contain words in the selected topics; and

a chunk modifier operable to modify the identified chunks.

24. Apparatus according to claim 23, wherein the chunk modifier is arranged to modify chunks by replacing them by ellipsis.

25. Apparatus according to claim 23, wherein the chunk modifier is arranged to modify chunks by causing them to be displayed so as to place less emphasis on the modified chunks.

26. Apparatus according to claim 25, wherein the chunk modifier is arranged to modify chunks to cause, when the outputter provides output data for display by a display, the modified chunks to be displayed using at least one of a smaller font size, a different font,

a different font characteristic and a different font colour from the other chunks.

27. Apparatus according to claim 23, wherein the chunk modifier is arranged to remove the identified chunks.

28. Apparatus according to claim 27, further comprising a processor operable to carry out syntactic or semantic processing on sentences from which chunks have been removed to maintain sentence coherence or cohesion.

29. Apparatus according to claim 23, wherein the chunk identifier is arranged to identify chunks by using punctuation marks to define the bounds of the chunks.

30. Apparatus according to claim 18, wherein the summary provider comprises a locater operable to locate words present in or representative of the content of the document data that co-occur with words in the topics; and the outputter is arranged to output summary data in which the or each topic is associated

with subsidiary items comprising located co-occurring words.

31. Apparatus according to claim 30, wherein the summary provider further comprises a further locater operable to locate all words present in or representative of the content of the document data that co-occur with the subsidiary items and the outputter is arranged to associate each such co-occurring word with the corresponding subsidiary item in the summary data.

32. Apparatus according to claim 31, wherein the summary provider further comprises a filter operable to filter the co-occurring words to select for the summary data those co-occurring words that themselves have co-occurrences with the subsidiary items.

33. Apparatus according to claim 1, further comprising a concept identifier operable to identify from the document data concepts that determine words representative of the content of the document data.

34. Apparatus according to claim 33, wherein the concept identifier is arranged to identify as

concepts at least one of synonyms, hypernyms and hypomyms in or relating to the document data.

35. Apparatus according to claim 33, wherein the concept identifier is arranged to access a lexical database to identify as concepts at least one of synonyms, hypernyms and hypomyms in or relating to the document data.

36. Co-occurrence significance calculating apparatus for use in text summarisation apparatus, the co-occurrence significance calculating apparatus comprising:

a co-occurrence identifier operable to identify as co-occurrences particular combinations of categories of words present in or representative of the content of document data;

a significance calculator operable to calculate a significance measure for the identified co-occurrences to determine significant ones of the identified co-occurrence; and

an outputter operable to output data representing the determined significant co-occurrences.

37. Apparatus according to claim 36, wherein the co-occurrence identifier is arranged to identify as co-occurrences at least some of the following combinations: noun and verb; noun and noun; noun and proper noun; verb and proper noun; and proper noun, and proper noun, and the significance calculator is operable to calculate a significance measure for the identified co-occurrences.

38. Apparatus according to claim 36, wherein the co-occurrence determiner is arranged to ignore the order of the words in the word combinations.

39. Apparatus for searching document data, the apparatus comprising:

a receiver operable to receive query terms supplied by a user;

a co-occurrence determiner operable to identify, for each query term, co-occurrences of words present in or representative of the content of the document data that include the query terms; and

an outputter operable to output parts or portions of the document data containing the identified co-occurrences.

40. Apparatus according to claim 39, wherein the co-occurrence determiner is arranged to identify as co-occurrences word combinations comprising words in particular grammatical categories.

41. Apparatus according to claim 39, wherein the co-occurrence determiner is arranged to identify as co-occurrences at least some of the following combinations: noun and verb; noun and noun; noun and proper noun; verb and proper noun; and proper noun and proper noun.

42. Apparatus according to claim 39, wherein the co-occurrence determiner is arranged to ignore the order of the words in the word combinations.

43. Apparatus for classifying topics in document data, which apparatus comprises:

a text splitter operable to split the document data into text segments;

a classifier operable to classify topics in the document data according to the distribution of the topics in the text segments so as to define main and subsidiary topics in the document data; and

an outputter operable to output data representing the classified topics.

44. Apparatus according to claim 43, wherein the classifier is arranged to determine that a topic is a main topic if the topic occurs in a predetermined percentage of the text segments and to classify any topic not meeting this requirement as a subsidiary or lesser topic.

45. Apparatus according to claim 43, wherein the classifier is arranged to weight a topic in accordance with the position in the document data of the text segment containing the topic.

46. Apparatus according to claim 43, wherein the classifier is arranged to weight a topic in accordance with the position in the document data of the text segment containing the topic so that a topic occurring in at least one of the first and last text segments of document data representing a document is given a higher weighting than topics occurring in the other text segments.

47. Apparatus for selecting sentences for use in a summary, the apparatus comprising:

a topic weight assigner operable to assign weights to topics in document data to be summarised;

a sentence weight assigner operable to assign weights to sentences in the document data;

a scorer operable to score each sentence in the document data by summing the assigned weights;

a selector operable to select the sentence or sentences having the highest score;

a topic weight adjuster operable to relatively reduce the weight allocated to topics in the selected sentence or sentences; and

a controller operable to cause the scorer, selector and topic weight adjuster to repeat the above operations until a certain number of sentences has been selected for the summary from the document data.

48. Apparatus according to claim 47, wherein the topic weight adjuster is arranged to set to zero the weight of any topic in the selected sentence or sentences.

49. Apparatus for providing a summary of document data, which apparatus comprises:

a receiver operable to receive data representing the topic or topics of the document data;

a locator operable to locate, for words in the or each topic, words in or representative of the content of the document data that co-occur with those words; and

an outputter operable to output summary data in which the or each topic is associated with subsidiary items comprising located co-occurring words.

50. Apparatus according to claim 49, wherein the summary provider further comprises a further locator operable to locate all words present in or representative of the content of the document data that co-occur with the subsidiary items and the outputter is arranged to associate each such co-occurring word with the corresponding subsidiary item in the summary data.

51. Apparatus according to claim 49, wherein the summary provider further comprises a filter operable to filter the co-occurring words to select for the summary data those co-occurring words that themselves have co-occurrences with the subsidiary items.

52. Apparatus for modifying chunks of sentences selected for a document data summary, which apparatus comprises:

a chunk identifier operable to identify chunks that do not contain words in topics representative of the content of the document data;

a chunk modifier operable to modify the identified chunks; and

an outputter operable to output the document data summary with the identified chunks of the selected sentences modified by the chunk modifier.

53. Apparatus according to claim 52, wherein the chunk modifier is arranged to modify chunks by replacing them by ellipsis.

54. Apparatus according to claim 52, wherein the chunk modifier is arranged to modify chunks by causing them to be displayed so as to place less emphasis on the modified chunks.

55. Apparatus according to claim 52, wherein the chunk modifier is arranged to modify chunks to cause, when the outputter provides output data for display by a display, the modified chunks to be displayed using

at least one of a smaller font size, a different font, a different font characteristic and a different font colour from the other chunks.

56. Apparatus according to claim 52, wherein the chunk modifier is arranged to remove the identified chunks.

57. Apparatus according to claim 56, further comprising a processor operable to carry out syntactic or semantic processing on sentences from which chunks have been removed to maintain sentence coherence or cohesion.

58. Apparatus according to claim 52, wherein the chunk identifier is arranged to identify chunks by using punctuation marks to define the bounds of the chunks.

59. Apparatus according to claim 52, further comprising a sentence selector operable to select the sentences for use in the summary data.

60. Apparatus according to claim 59, wherein the sentence selector comprises:

a topic weight assigner operable to assign weights to the topics;

a sentence weight assigner operable to assign weights to sentences in the document data;

a scorer operable to score the sentences by summing the assigned topic and sentence weights; and

a selector operable to select the sentence or sentences having the highest score or scores for the summary.

61. Apparatus according to claim 52, wherein the sentence selector comprises:

a topic weight assigner operable to assign weights to the topics;

a sentence weight assigner operable to assign weights to sentences in the document data;

a scorer operable to score the sentences by summing the assigned topic and sentence weights;

a selector operable to select the sentence or sentences having the highest score or scores;

a topic weight adjuster operable to reduce the weight allocated to the topic or topics in the selected sentence or sentences; and

a controller operable to cause the scorer, selector and topic weight adjuster to repeat the above

operations until a predetermined number of sentences has been selected for the summary from the document data.

62. A method of identifying topics of document data, the method comprising a processor carrying out the steps of:

ranking words that are present in or representative of the content of the document data;

ranking co-occurrences of words that are present in or representative of the content of the document data;

ranking phrases in the document data;

selecting the highest ranking words;

identifying which of the highest ranking co-occurrences contain at least one of the highest ranking words;

identifying the phrases containing at least one word from the identified co-occurrences;

selecting the highest ranking one or ones of the identified phrases as the topic or topics of the document data; and

outputting data relating to the selected topics.

63. A method of calculating co-occurrence significances for use in text summarisation apparatus, the method comprising a processor carrying out the steps of:

identifying as co-occurrences particular combinations of categories of words present in or representative of the content of document data;

calculating a significance measure for the identified co-occurrences to determine significant ones of the identified co-occurrence; and

outputting data representing the determined significant co-occurrences.

64. A method of searching document data, the method comprising a processor carrying out the steps of:

receiving query terms supplied by a user;

identifying, for each query term, co-occurrences of words present in or representative of the content of the document data that include the query terms; and

outputting parts or portions of the document data containing the identified co-occurrences.

65. A method of classifying topics in document data, which apparatus comprises a processor carrying out the steps of:

splitting the document data into text segments;

classifying topics in the document data according to the distribution of the topics in the text segments so as to define main and subsidiary topics in the document data; and

outputting data representing the classified topics.

66. A method of for selecting sentences for use in a summary, the method comprising a processor carrying out the steps of:

assigning weights to topics in document data to be summarised;

assigning weights to sentences in the document data;

scoring each sentence in the document data by summing the assigned weights;

selecting the sentence or sentences having the highest score;

relatively reducing the weight allocated to topics in the selected sentence or sentences; and

repeating the scoring, selecting and topic weight adjusting steps until a certain number of sentences has been selected for the summary from the document data.

67. A method of providing a summary of document data, which method comprises a processor carrying out the steps of:

receiving data representing the topic or topics of the document data;

locating, for words in the or each topic, words in or representative of the content of the document data that co-occur with those words; and

outputting summary data in which the or each topic is associated with subsidiary items comprising located co-occurring words.

68. A method of modifying chunks of sentences selected for a document data summary, which method comprises a processor carrying out the steps of:

identifying chunks that do not contain words in topics representative of the content of the document data;

modifying the identified chunks; and

outputting the document data summary with the modified identified chunks of the selected sentences.

69. Program instructions for programming a processor to carry out a method in accordance with claim 62.

70. A storage medium storing program instructions in accordance with claim 69.

71. A signal carrying program instructions in accordance with claim 69.

72. Apparatus for identifying topics of document data, the apparatus comprising:

word ranking means for ranking words that are present in or representative of the content of the document data;

co-occurrence ranking means for ranking co-occurrences of words that are present in or representative of the content of the document data;

phrase ranking means for ranking phrases in the document data;

words selecting means for selecting the highest ranking words;

co-occurrence identifying means for identifying which of the highest ranking co-occurrences contain at least one of the highest ranking words;

phrase identifying means for identifying the phrases containing at least one word from the identified co-occurrences;

phrase selecting means for selecting the highest ranking one or ones of the identified phrases as the topic or topics of the document data; and

output means for outputting data relating to the selected topics.